

Prediction of Fujisaki Model's Phrase Commands

João Paulo Teixeira*, Diamantino Freitas** and Hiroya Fujisaki***

{*Polytechnic Institute of Bragança, **Faculty of Engineering of University of Porto}, Portugal

***University of Tokyo, Japan

joaopt@ipb.pt, dfreitas@fe.up.pt, fujisaki@alum.mit.edu

Abstract

This paper presents a model to predict the phrase commands of the Fujisaki Model for F0 contour for the Portuguese Language. Phrase commands location in text is governed by a set of weighted rules. The amplitude (Ap) and timing (T0) of the phrase commands are predicted in separate neural networks. The features for both neural networks are discussed. Finally a comparison between target and predicted values is presented.

1. Introduction

This paper reports the work towards a part of the prosody model for TTS that is under development for the European Portuguese language.

F0 is the most perceptually relevant component in prosody. The Fujisaki model of F0 was been proven to be well adapted with very high naturalness to several languages [1] like Japanese, Korean, Spanish, Polish, Greek, Swedish, English [2], German [3], Basque [4] and now also Portuguese. The Fujisaki model [1] consists of the logarithmic addition of baseline fundamental frequency, phrase components and accent components. The baseline fundamental frequency is constant in an utterance. The phrase components are parameterized with phrase commands (CF) as a set of impulses and the accent components with accent commands (CA) as a set of pedestal functions.

This paper is dedicated to describe the prediction of only the CF. The prediction of the CAs will be done using the information of CF and other features in subsequent studies. Alpha, the natural angular frequency parameter of the CF, is considered to be equal to 2.0 for all CF. This value allows the best fit to the F0 contour for the presented database. The task is to predict from the text the position where the CFs will be inserted, as well as each of the amplitudes Ap, and the distances of each CF to the associated time position, T0, in the speech data stream. The prediction of these parameters will be presented in the next sections.

The text was structured in the following units: paragraphs, sentences, phrases, and accent groups. The beginnings of accent groups were considered the eligible positions for placement of a CF.

2. Text and Speech Corpus and F0 labelling

The data used for training and testing were extracted from the FEUP-IPB database [5]. This database is centred on several texts extracted from newspapers that were read by a male professional radio broadcast speaker at the average speech rate 12.2 phonemes/second. The speech waveforms were then manually labelled in three levels: the phonetic level, considering 46 different segments classes and also marking the tonic syllable; the word level, marking beginning and

ending of words; and, thirdly, the phrase level, marking beginning and ending of phrases as well as all orthographic punctuation marks. Seven texts of the data-base were used, in a total of 101 paragraphs of high variety dimensions, from one to one hundred words. Mainly declarative and interrogative types of sentences were selected, in a total of 18.700 segments in 21 minutes of speech. The corpus was divided into two sets, a training set consisting of about 80% of the paragraphs and the test set with the remaining 20%.

The Fujisaki parameters were extracted using a specifically developed tool. The process of labeling starts with an automatic CF prediction algorithm developed by Mixdorff [6], and was followed by a manual optimization of CF and introduction of, as a rule, one CA for each syllable that has its own F0 movement. The optimization was oriented to the best fit of the model predicted contour to the original F0 in voiced parts. The manual optimization allowed the improvement not only of the root-mean-squared error (rmse) calculated along the corpus between the two F0 contours in voiced parts, but also of the naturalness of the re-synthesized speech with the model F0.

3. CF estimation from text

The estimation of CF from text addresses two issues. The first is to determine its insertion position in text, and the second is to estimate the amplitude (Ap) and the distance (T0) to the time point in speech associated with the text position. These two issues are solved in separate steps described below.

Table 1: Numbers of punctuation marks, associated CFs and percentages of coincidence

Orthographic punctuation	# of occurrence	# of CF	%
.	67	64	96
,	379	261	69
?	12	12	100
!	4	3	75
...	1	1	100
-	7	6	86
;	2	2	100
:	6	5	83

3.1. CF positions in text

From the analysis of the location of CF it is quite obvious that some orthographic punctuation marks impose presence of a CF. Table 1 presents the percentage of occurrences of orthographic punctuations that originate CFs. In this table the punctuation marks at end of paragraph are excluded, due to the obvious impossibility of being associated with CF. Although punctuation marks “!” “...” “-” “;” “:” do not present

statistical relevance, the table suggests to have one CF associated to each orthographic punctuation mark. In case of comma “,” the percentage is not higher basically due to the proximity of some comas to other punctuation marks.

Besides the CFs imposed by orthographic marks, there are other CF, about 30% of total, not linked with the punctuation.

The algorithm of the tool described above, in 2, to govern the location of insertion of CF, will, in the first step, place CFs linked to orthographic punctuation, and subsequently consider several candidate positions to insert other CFs. For each candidate position the score will be calculated by a mathematical model as described in next section.

3.1.1. CFs not linked with orthographic marks

This section deal only with this type of CF. The objective is to find anchors to associate them. Text speech analysis of several of these CF, suggests that different factors seems to contribute to their locations. Factors like distance to previous CF, distance to next CF, presence of pause, length of previous word and type of next word, were statistically analyzed and correlated with the presence of this type of CF.

For each candidate position, one score will be calculated by expression (1) that combines the weights of each factor.

$$S = W_{pCF} \times W_{nCF} \times (W_p + W_{lpw} + W_{tw}) \quad (1)$$

In equation (1), S is the score for the candidate position, W_{pCF} , W_{nCF} , W_p , W_{lpw} , and W_{tw} are the weights for previous CF distance, next CF distance, pause, length of previous word and type of next word, respectively.

The distances to previous and next CF factors have different histograms, but both are near to a normal distribution. Table 2 presents the relevant statistical data. Weights for previous and next candidate are given by the normal probability density function with the respective mean and standard deviation presented in Table 2 at the respective distance to previous and next CF.

Table 2: Statistical data of distance to previous and next CF

Statistical data	Distance to previous CF (s)	Distance to next CF (s)
Minimum	0.55	0.75
Maximum	3.26	4.77
Mean	1.70	1.94
Standard deviation	0.53	0.65

The weight for presence of pause, W_p , is 1 or 0 in case of presence or not of pause.

For the weight relative to the length of the previous word, W_{lpw} , the length of the word is considered plus the length of the eventual pause. This factor assumes a higher correlation with presence of CF, for values above 0.5 s. The weight used for this factor is given by equation (2).

$$W_{lpw} = \log(5 \times (\text{length} + 0.2)) \quad (2)$$

The weight, W_{tw} for the type of next word, was determined according to the correlation of some words with this type of CF and is given by a table. This table, containing the most correlated words, has weights between 0.7 and 1. For other words not in table W_{tw} is 0.7, 0.5 and 0.2 for words with one two or more syllables.

3.1.2. Algorithm to insert CF

The eligible positions to inserted CFs are only the position of the beginning of the accent groups. The exact time position will be determined by subtracting the T0 value predicted by the neural network (NN) as will be described in 3.2. The algorithm was designed from several observations of the positions of CFs.

The algorithm starts by inserting CFs just after the punctuation marks of Table 1. Then it removes the CF whose distance to the previous is less than 1 s if previous sentence is not of interrogative type.

Then, for the intervals between CFs that are longer then 3 s, it identifies candidate positions to insert a new CF. The candidate positions are the start positions of accent groups between previous CFs plus 0.6 s, and the minimum between next CF minus 0.75 s and previous CF plus 3.25 s. These limits for candidates ensure the minimum distance to previous and next CF, according to Table 2. Then it calculates the score S for each candidate according to (1), and only the maximum score candidate will be considered. If the maximum score candidate was a score greater than 1, then one CF is inserted in its position. The process is repeated with the new set of CF until the end of the paragraph.

3.1.3. Evaluation of inserted CF

A comparison between the positions of labeled CF and the ones inserted by the algorithm is given in Tables 3 and 4. The position of labeled CF is the reference position meanwhile the inserted CF position will be affected by T0. Table 3 shows the number of inserted and labeled CFs, which are very close, as well as the average and standard deviation of respective distances. The histograms of distances of labeled and inserted CFs are quite similar in shape. Table 4, presents the number of rightly inserted CFs determined as the number of inserted CFs at position less distant than X seconds from the nearest labeled CF, and the number of wrongly inserted CFs as the number of inserted CFs whose distance to the nearest labeled CF is longer then X, and not inserted CF as the number of labeled CFs without inserted CF at distance X or less. The range X is a tolerance for the T0 that will affect the exact position of inserted CF. The maximum T0 is almost 1 s.

Visual inspection indicates that the inserted CFs are generally in a coherent position.

3.2. Prediction of Ap and T0 parameters

The amplitude, A_p , and distance, T0, of the CF to the beginning of the accent group, are predicted in a second step by means of a neural network. Because of the low correlation (0.081) between A_p and T0, one neural network was developed for each parameter. The performances for both parameters are improved by using the two NNs instead of one NN for both parameters.

Several architectures were considered and tested for both NNs. The selected architecture to predict A_p , is a feed-forward type with two hidden layers with two nodes each and with hyperbolic logarithmic activating functions. The output layer is one node with a linear activating function. The output is 85% of A_p divided by the maximum A_p and normalized to have null average and standard deviation equal to 1. The training algorithm was a back-propagation Levenberg-Marquardt [7].

Table 3: Comparison distance between labeled and inserted CFs. The number of CFs, the minimum, maximum and average distances and standard deviations in seconds

CF	#	Dist_mn	Dist_mx	Averg.	Std.
Labeled	646	0.55	4.78	1.86	0.66
Inserted	643	0.50	2.99	1.83	0.48

Table 4: Number of rightly inserted CFs (ri), wrong inserted (wi) and not inserted (ni) at a tolerance time distance X.

	X=0.6 s	X=0.8 s	X=1 s
ri	494	570	604
wi	149	73	39
ni	158	91	71

The architecture of the T0 NN is also a feed forward type with two hidden layers, but with four nodes in first hidden layer and a hyperbolic tangent activating function, and two nodes in the second hidden layer activated by a hyperbolic logarithmic function. The output layer is also one node with a linear activating function. The output is T0, preprocessed as Ap. This NN was trained with the same algorithm.

3.2.1. Text and speech features for Ap and T0

We now introduce the features that were extracted in order to evaluate its relevance. Each feature was coded in one input node of the NN.

Although some features presented below in Table 5 don't have an individually significant correlation with T0, suggesting exclusion from the NN, in fact, all together, their presence, improves the prediction performance.

Some features are highly mutually correlated as is the case of features 3 and 6, 4 and 7, 11, 12 and 13, 15 and 16, and finally, 18 and 19. Anyhow they don't carry exactly the same information, and their ensemble use improves the performance. An explanation of the features, as measured in Table 5, follows:

1. the correlation between most of the marks presented in Table 1 and Ap are similar, and not relevant with T0. Therefore just the comma and the full stop were classified separately. This feature was coded in four levels: other mark, full stop, comma, no mark;
2. only the interrogative type of sentence showed a different correlation with Ap and T0. Therefore this feature was coded in the levels of interrogative type or other type. Different types of interrogatives were not distinguished;
3. correlation with Ap indicates higher Ap in the beginning of sentences.
4. correlation indicates lower Ap in the end of sentences;
5. correlation shows lower Ap for long sentences;
6. correlation indicates higher Ap in the beginning of paragraph;
7. correlation indicates lower Ap in the end of paragraph;
8. correlation indicates higher Ap in the first sentences of paragraph;
9. correlation indicates lower Ap in the last sentences of paragraph;

Table 5: Set of features and their correlations r with Ap and T0

F #	Feature description	r(F,Ap)	r(F,T0)
1	Orthographic mark	-0.470	0.010
2	Interrogative sentence	0.075	0.330
3	Index # of CF in sentence, from beginning	-0.380	0.025
4	Index # of CF in sentence, from end	0.177	0.041
5	Length of sentence (s)	-0.127	0.051
6	Index # of CF in paragraph from beginning	-0.448	0.026
7	Index # of CF in paragraph from end	0.239	0.027
8	Index # of sentence in paragraph from beginning	-0.185	0.020
9	Index # of sentence in paragraph from end	0.117	0.008
10	Length of preceding pause (s)	0.569	0.067
11	CF in beg. position of phrase	0.223	-0.017
12	CF in beg. position of sentence	0.460	0.025
13	CF in beg. pos. of paragraph	0.572	0.030
14	Tonic syllable in the beginning of the accent group	0.052	0.074
15	Distance to the preceding CF (s)	0.534	0.213
16	Distance in syllables to the preceding CF	0.525	0.126
17	Orthographic mark of the preceding CF	0.279	0.032
18	Distance to the next CF (s)	0.221	-0.323
19	Distance in syllables to the next CF	0.241	-0.285
20	Orthographic mark of the next CF	0.140	0.003

10. is the length of pause if there is one just before de CF. This feature is highly correlated with Ap;
11. indication if the CF is in beginning position of a phrase. This position is correlated with higher Ap;
12. indication if the CF is in beginning position of a sentence. This position is correlated with higher Ap;
13. indication if the CF is in beginning position of a paragraph. This position is correlated with higher Ap;
14. indication if the accent group starts with a tonic syllable. Slightly correlated with higher Ap and longer T0;
15. highly correlated with Ap and T0;
16. highly correlated with Ap and T0;
17. similar with feature 1, but coded in different order due to different levels of correlation: other mark, coma, no mark, full stop;
18. is the CF length. As longer is the CF length, higher is the Ap and shorter the T0. Is the most relevant feature for T0;
19. similar correlations with previous feature;
20. in this feature others marks are relevant. Therefore, it is coded in the following six levels: "other mark", ",", "...", ";", "?", "no mark".

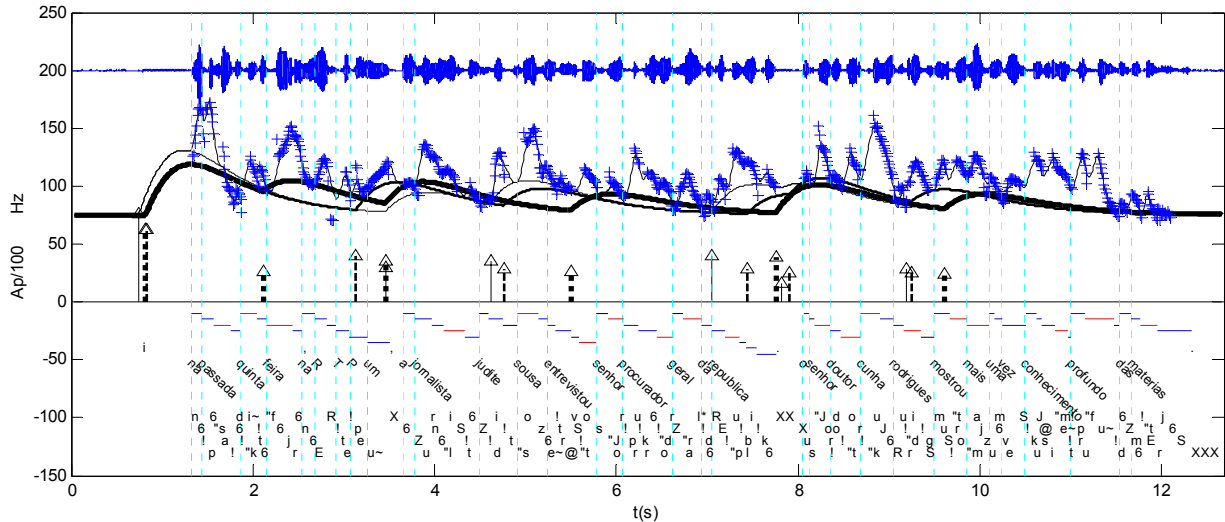


Figure 1: Inserted CFs in Paragraph: Na passada quinta-feira, na RTP1 a jornalista Judite de Sousa entrevistou o senhor Procurador geral da República. O Senhor Doutor Cunha Rodrigues mostrou mais uma vez conhecimento profundo das matérias (Last Thursday, in RTP 1, the journalist Judite de Sousa, interviewed mister *procurador geral da República*. Mister Doctor Cunha Rodrigues showed once again a deep knowledge of matters). From top to bottom: speech waveform; + signs- measured F0; *thin line* – phrase components and phrase components plus accent components direct from labeled parameters; *medium line* – phrase components from predicted CFs Ap and T0, considering initial positions of labeled CF; *thick line* – phrase components from totally predicted CFs; CFs: *thin line* – labeled, *dashed medium line* – Ap and T0 predicted with NN considering initial positions of labeled CF, *dotted thick line* – predicted CFs; lines identifying syllables (the accent groups are the groups of syllables in descending order); orthographic marks in text; words; phoneme labels.

3.2.2. Training of the Neural Networks

Training was done over the mentioned training set and using the test set for cross-validation in order to avoid over-fitting. The test vector was used to stop training early if further training on the training set will hurt generalization to the test set. The cost function used for training was the mean squared error between output and target values.

3.2.3. Testing

The best linear correlation coefficients (r) between predicted and measured Ap and T0 obtained for the test set are **0.772** and **0.646**, respectively.

Fig. 2 presents the predicted CFs for one example paragraph. The T0 and Ap predicted from initial positions of labeled CF, depicted with medium line in the figure, show the output of both NNs. The thick line represents the result of prediction of CFs from text. This line presents in our opinion a reasonable phrase component for this paragraph.

4. Discussion and Conclusions

The current study presents a model to predict CFs of the Fujisaki model from text in Portuguese. The model performs in two steps. In the first step it inserts CFs associated with the beginning of accent groups, based on orthographic marks and weighted candidates. The second step predicts the exact position with T0 and Amplitude Ap with two NNs.

The location of inserted CFs seems to be consistent with text and with labeled CFs. The best linear correlation coefficient of the prediction of Ap and T0 are 0.772 and 0.646, respectively. A global final evaluation of the model, with perceptual tests, can be done only when it will be

complete. Meanwhile, from the analysis of several paragraphs, the phrase components calculated from the predicted CFs, we conclude that with a good set of CAs the resulting F0 contour fits the original one with a good closeness. The task of predicting accent commands is now under development. Anyhow, as can be seen in Figure 1 the application of Fujisaki model to predict F0 in Portuguese is very promising.

5. References

- [1] Fujisaki, H., "Modeling in study of Tonal Features of Speech with Application to Multilingual Speech Synthesis", Proc. of Joint International Conference of SNLP and Oriental COCODA. May 2002, Thailand.
- [2] Fujisaki, H., Ohno, S., "Analysis and modeling of fundamental frequency contours of English Utterances", in *Eurospeech '95, Madrid*.
- [3] Mixdorff, H., "An Integrated Approach to Modeling German Prosody", Thesis for Dr.-Ing. Habil., Technical University of Dresden. 2002.
- [4] Navas, E., Hernandez, I., Sanchez, J. M., "Basque Intonation Modelling for Text to Speech Conversion", in *Proceedings of ICSLP '02, Denver*.
- [5] Teixeira, J. P., Freitas, D., Braga, D., Barros, M. J., Latsch, V., "Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB", in *Eurospeech '01, Aalborg*.
- [6] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", in *Proceedings ICASSP 2000, vol. 3, 1281-1284, Istanbul, Turkey*.
- [7] Hagan, M. T., Menhaj, M., "Training feedforward networks with the Marquardt algorithm", *IEEE Transactions on Neural Networks*, vol. 5, n 6, 1994.